

## **Abstract Title Page**

**Title:** Synthesizing results from replication studies using robust variance estimation: Corrections when the number of studies is small

**Authors and Affiliations:**

Elizabeth Tipton, *Teacher College, Columbia University*

## **Abstract Body.**

### **Background / Context:**

Replication studies allow us to make comparisons and generalizations regarding the effectiveness of an intervention across different populations, versions of a treatment, settings and contexts, and outcomes. One method for making these comparisons across many replication studies is through the use of meta-analysis. Meta-analysis methods allow us to answer questions like: On average, how effective are interventions of this type? How much does the effectiveness vary across studies? And, often most importantly, does the effectiveness vary in relation to features of the underlying populations, treatments, settings, or outcomes?

In many experiments, the effectiveness of a treatment is measured using multiple outcomes. For example, in reading intervention studies, measures of fluency, word recognition, and comprehension might be collected. In some studies, questions of durability of a treatment effect are important; in order to assess this, measures might be collected both at the end of an intervention and three- or six-months later. Traditional meta-analytic methods, however, have required effect sizes to be independent, making it difficult for inferences and comparisons to be made across different outcomes. In order to ensure independence, the common solution is to either select only one outcome from, or to create a single combined measure for, each study for inclusion in the meta-analysis. This often results in a loss of information.

A recent innovation in meta-analysis is the introduction of a robust variance estimator that allows for the inclusion of multiple, correlated effect sizes in a meta-analysis (Hedges, Tipton, and Johnson, 2010). This method does not require any information on the true correlation structure of these estimates, which is particularly important since this information is rarely available in primary studies. The statistical theory behind the robust variance estimation (RVE) method is asymptotic; in large-samples, it has been shown to be an unbiased estimator of the true sampling variance. The RVE approach is already widely used in meta-analyses in psychology, social welfare, and education.

Importantly, the RVE estimator is a type of linearization or Taylor-series estimator, which are commonly used in the analysis of panel data in econometrics, in survey sampling (with complex sampling designs), with generalized estimating equations, and are particularly useful when a standard regression model is preferred and the random effects are not of direct interest. In particular, the RVE approach is most similar to that of *clustered standard errors* (Liang & Zeger, 1986) which are used to account for the clustering or nesting of data (e.g., students in schools); clustered standard errors are an extension to *Huber-White standard errors* (Huber 1967; White, 1980), which are used for accounting for heteroskedastic errors in independent data.

### **Purpose / Objective / Research Question / Focus of Study:**

While the RVE estimator is unbiased in large-samples, its small sample properties are often not ideal. Previous simulation studies have shown that over many different effect sizes, when the number of studies is less than 40, the associated confidence intervals often under-cover and the associated hypothesis tests have Type I error rates far above nominal (Hedges, et al, 2010; Tipton, 2013; Williams, 2012). While these studies have varied the effect size and number of primary studies, however, other conditions – including the number and types of covariates used in the meta-regression models – have not been studied. To date, the main conclusion from these studies is that RVE results should not be trusted with meta-regression models with fewer

than 40 studies. This is a real limitation for the method given that at 50% of meta-analyses in education contained fewer than 40 studies (Ahn, Ames, & Myers, 2012).

### Significance / Novelty of study:

This paper investigates possible approaches to adjusting the RVE estimator when the number of studies is small (less than 40), which is common in the both meta-analyses and replication studies in education. These adjustments are based on work by Bell and McCaffrey (2002) and McCaffrey, Bell, and Botts (2001), which themselves are extensions to adjustments found in MacKinnon and White (1985). These include three methods for adjusting the residuals used in RVE and two methods for adjusting the degrees of freedom used for making inferences. In order to evaluate how well these methods perform in practice, we present results of two simulation studies: in the first study, we focus on several meta-regression models with a single covariate, while the second study focuses on a larger meta-regression model that mirrors the type of models found in practice.

### Statistical, Measurement, or Econometric Model:

The RVE approach can be used whenever researchers seek to combine information across studies and at least one of these studies includes multiple outcomes. The fundamental problem with combining these effect sizes is that they are not independent. There are two types of correlation structures addressed by this method: “correlated effects”, which arise from multiple measures on the same units, and “hierarchical effects”, which occur because independent experiments conducted in the same laboratory often share many features (e.g., protocols, study populations).

RVE can be used to estimate both an average effect size across all studies (and outcomes) and for estimating meta-regression models. These models allow for comparisons to be made across features of the population, versions of the treatment, types of outcomes, and features of the study context or setting.

Let study  $j = 1 \dots m$  have a vector of  $k_j$  effect size estimates  $\mathbf{T}_j$ , a design matrix  $\mathbf{X}_j$ , and a weight matrix  $\mathbf{W}_j$ . Here  $\mathbf{X}_j$  arises from the design of the meta-analysis and may include an intercept as well as covariates that vary across studies or effect sizes. Assume each study  $j=1 \dots m$  also has an associated vector of  $k_j$  residuals  $\boldsymbol{\epsilon}_j$ . We can relate these via the regression

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where  $\mathbf{T} = (\mathbf{T}'_1, \dots, \mathbf{T}'_m)'$  is a vector of  $m$  vectors, each with  $k_j$  effect size estimates,  $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_m)'$  is a design matrix of  $m$  stacked matrices, each of dimension  $k_j \times p$ , and  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of coefficients to be estimated. Finally, let  $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_1, \dots, \boldsymbol{\epsilon}'_m)'$  be the vector of stacked error vectors, each of dimension  $k_j \times 1$ .

The regression coefficients  $\boldsymbol{\beta}$  can be estimated using weighted least squares as

$$\mathbf{b} = \left( \sum_{j=1}^m \mathbf{X}'_j \mathbf{W}_j \mathbf{X}_j \right)^{-1} \left( \sum_{j=1}^m \mathbf{X}'_j \mathbf{W}_j \mathbf{T}_j \right)$$

The robust variance estimation method proposes to estimate the  $V(\mathbf{b})$  using

$$\mathbf{V}^R = \left( \sum_{j=1}^m \mathbf{X}'_j \mathbf{W}_j \mathbf{X}_j \right)^{-1} \left( \sum_{j=1}^m \mathbf{X}'_j \mathbf{W}_j \mathbf{A}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{A}_j' \mathbf{W}_j \mathbf{X}_j \right) \left( \sum_{j=1}^m \mathbf{X}'_j \mathbf{W}_j \mathbf{X}_j \right)^{-1} \quad (1)$$

where  $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j \mathbf{b}$  is the  $k_j \times 1$  residual vector in the  $j^{\text{th}}$  study and, in the standard RVE estimator,  $\mathbf{A}_j = \mathbf{I}_{k_j}$ . Based on this, hypotheses of the form  $\beta_k = 0$  can be tested using the Wald statistic

$$t_k^R = \frac{b_k}{\sqrt{V_{kk}^R}}. \quad (2)$$

In order to test if  $\beta_k = 0$ , this test rejects the null hypothesis if  $|t_k^R| \geq t_{m-p, \alpha}$ , where  $t_{m-p, \alpha}$  is the level- $\alpha$  t-value with  $df_{HTJ} = m - p$  degrees of freedom.

In this paper, we investigate two small sample corrections, one to the residuals and another to the degrees of freedom. The corrections to the residuals come through the  $\mathbf{A}_j$  adjustment matrix shown in Eqn. (1) above. We investigate three such adjustments:

- 1)  $\mathbf{A}_j^{\text{HTJ}} = [m/(m-p)]^{1/2} \mathbf{I}_{k_j}$ .
- 2)  $\mathbf{A}_j^{\text{JK}} = [m/(m-1)]^{1/2} (\mathbf{I}_{k_j} - \mathbf{H}_{jj})^{-1}$ , and
- 3)  $\mathbf{A}_j^{\text{MBB}} = (\mathbf{I} - \mathbf{H}_{jj})^{-1/2}$

where  $\mathbf{H}_{jj} = \mathbf{X}_j \mathbf{Q} \mathbf{X}_j' \mathbf{W}_j$  and  $\mathbf{Q} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$ . Second, we investigate adjustments to the degrees of freedom. The first correction, proposed by Hedges et al (2010), is to use the t-distribution with  $df_{HTJ} = m - p$ , where  $m$  is the number of studies and  $p$  is the number of predictors in the meta-regression model. The second correction, proposed by McCaffrey, Bell, and Botts (2001) is to estimate the degrees of freedom using the Satterthwaite (1946) approximation. This results in

$$df_{sk} = (\sum \lambda_{jk})^2 / \sum \lambda_{jk}^2.$$

where  $\lambda_{jk}$  are the  $k_j$  eigenvalues of  $\Sigma^{1/2} (\sum \mathbf{g}_{jk} \mathbf{g}_{jk}') \Sigma^{1/2}$ , for the covariance matrix  $\Sigma = E(\epsilon \epsilon')$ , which is a block diagonal matrix composed of the  $m$   $\Sigma_j$  matrices. In the full paper, we provide specific estimation strategies for both the residual adjustments ( $\mathbf{A}_j$ ) and the degrees of freedom adjustments ( $df_{sk}$ ) particular to the weighting strategies and correlation problems found in RVE in meta-analysis (i.e., correlated effects, hierarchical effects).

The three residual corrections and the two degrees of freedom adjustments lead to a combination of 6 possible small sample corrections to RVE. In order to investigate how well these perform in small samples, we conducted two simulation studies. In the first simulation study we focus on simple meta-regression models, each with only one covariate. We focus here on the role of variable type on degrees of freedom, Type I error rates, and statistical power. We are interested in the role of variable type since previous research suggests that statistical properties of the  $V_{kk}^*$  estimators depend on both the degree to which the covariates are balanced and on the leverage of the observations (Bell & McCaffrey, 2002; Chesher & Austin, 1991; Long & Ervin, 2000; MacKinnon, 2013). In the second simulation study, we attempt to mirror practice more closely by comparing properties of the six corrections for meta-regression models with 4 covariates. Here we focus on  $m=20$  studies and include all four covariate types found in Study 1.

Both simulation studies focus on the test found in (2) above. For the  $\alpha$ -level  $\alpha=0.05$ , in both Study 1 and Study 2 we investigate how Type I error rates vary in relation to the number of studies ( $m$ ), and the types of predictors; in addition, in Study 1, we investigate the statistical power of the test for three true regression coefficient relationships (small, medium, large).

## Findings / Results:

The results of our simulation studies include 4 main findings:

- 1) The most important result is that the estimator proposed by Hedges et al (2010) only performs well in very limited circumstances; when covariates are unbalanced or have high

leverage, and particularly when the number of studies is small, the Type I error rates can be tremendously larger than the stated  $\alpha = 0.05$ . (See Figures 1 and 2).

2) The second major finding is that the largest improvements to RVE arise through the use of Satterthwaite degrees of freedom, and that no estimator performs well when these degrees of freedom are smaller than 4.

3) Third, our simulations suggest that two estimators perform well in a wide variety of situations: the bias reduced linearization estimator for weighted least squares proposed by McCaffrey, Bell, and Botts (MBBS; 2001) and the jackknife estimator (JKS). (See Figures 1 and 2).

4) The jackknife estimator (JKS) is typically more conservative than the MBBS estimator, and, as a result, it is also less powerful. Over the parameters included in our study of power, the MBBS is uniformly more powerful than the jackknife. (See Figure 3).

### **Usefulness / Applicability of Method:**

In order to illustrate the usefulness of the method, we include an example based on a meta-analysis by Tanner-Smith and Lipsey (2013). This meta-analysis combined results of randomized-experiments evaluating the effectiveness of brief alcohol interventions ( $< 5$  hours of contact time,  $< 4$  weeks in duration) among adolescents and young adults. In these analyses, the outcomes include measures of the first alcohol consumption after the experiment ended. In the example, we focus on a subset of  $m = 28$  studies (containing 300 effect sizes). Given the findings of the simulation studies, we compare results based on the original estimator given by Hedges et al (2010), and the MBBS and JKS estimators developed in this paper.

We focus on a meta-regression model with 4 covariates. These include a variety of variable types, similar to the types under study in our simulation studies. We estimated this model in the statistical program **R** (R Development Core Team, 2012) using a correlated effects RVE model with an assumed  $\rho = 0.80$ . The results of the meta-regression are presented in Table 1 in Appendix B. This table illustrates two points. First, that the (recommended) Satterthwaite degrees of freedom vary highly from covariate to covariate, and for some covariates (even with  $m = 28$  studies) can be quite small. Second, the p-values differ between the three tests, with the most liberal results coming from the unadjusted results (Hedges et al) and the most conservative from the jackknife (JKS); the MBBS results are in between.

### **Conclusions:**

As our example illustrates, using the MBB or JK estimators with Satterthwaite degrees of freedom can impact the conclusions drawn from a robust meta-regression. Most commonly these differences arise because of degrees of freedom differences. These degrees of freedom differences can be large and are directly related to the degree of balance and maximum leverage in the data. Importantly, they often lead to different conclusions.

Importantly, since the Satterthwaite degrees of freedom of these small sample adjustments depends not just on the number of studies, but also on the type of variable (dichotomous, continuous), the level of the covariate (study, effect size), the degree of balance across studies, and the presence of high leverage values, our simulation studies suggest that it is difficult to know at what point small sample corrections are no longer needed. Even with  $m = 40$  studies, the probability of a Type I error for the standard RVE estimator can be much larger than  $\alpha = 0.05$ . For this reason we argue that it is best if the corrections provided here are implemented in *all* RVE analyses, even those of moderate to large sizes.

## Appendices

### Appendix A. References

- Ahn, S., Ames, A.J., & Myers, N.D. (2012) A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82(4): 436-476.
- Bell, R.M. & McCaffrey, D.F. (2002) Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169-181.
- Chesher, A. & Austin, G. (1991) The finite-sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics*, 47: 153-173. Elsevier Science Publishers.
- Hedges, L.V., Tipton, E., & Johnson, M. (2010) Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1): 39-65. Erratum in 1(2): 164-165.
- Huber P. (1967) The behavior of maximum-likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. LeCam, L.M. & Neyman, J. University of California Press: Berkeley; pp. 221-233.
- Liang, K.L. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13-22.
- Long, J.S. & Ervin, L.H. (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3): 217-224
- MacKinnon, J.G. (2013) Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, eds. X. Chen and N.R. Swanson, pp. 437-461. New York: Springer.
- MacKinnon, J.G. & White, H. (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29:305-325.
- McCaffrey, D.F., Bell, R.M., & Botts, C.H. (2001) Generalizations of biased reduced linearization. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.
- R Development Core Team (2012) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Satterthwaite, F. (1946) An approximate distribution of estimates of variance components. *Biometrics*, 2: 110-114.
- Tanner-Smith, E. E., & Lipsey, M. W. (2013). A meta-analysis of brief intervention effects for adolescents and young adults.
- Tipton, E. (2013) Robust Variance Estimation in Meta-regression with Binary Dependent Effects. *Research Synthesis Methods*, 4(2): 169-187.
- White H. (1980) A heteroscedasticity-consistent covariance matrix and a direct test for heteroscedasticity. *Econometrica*, 48:817-838.
- Williams, R. (2012) Using robust standard errors to combine multiple regression estimates with meta-analysis. (Doctoral Dissertation) Retrieved from *ProQuest Dissertations and Theses* (Accession Order No. 3526372)

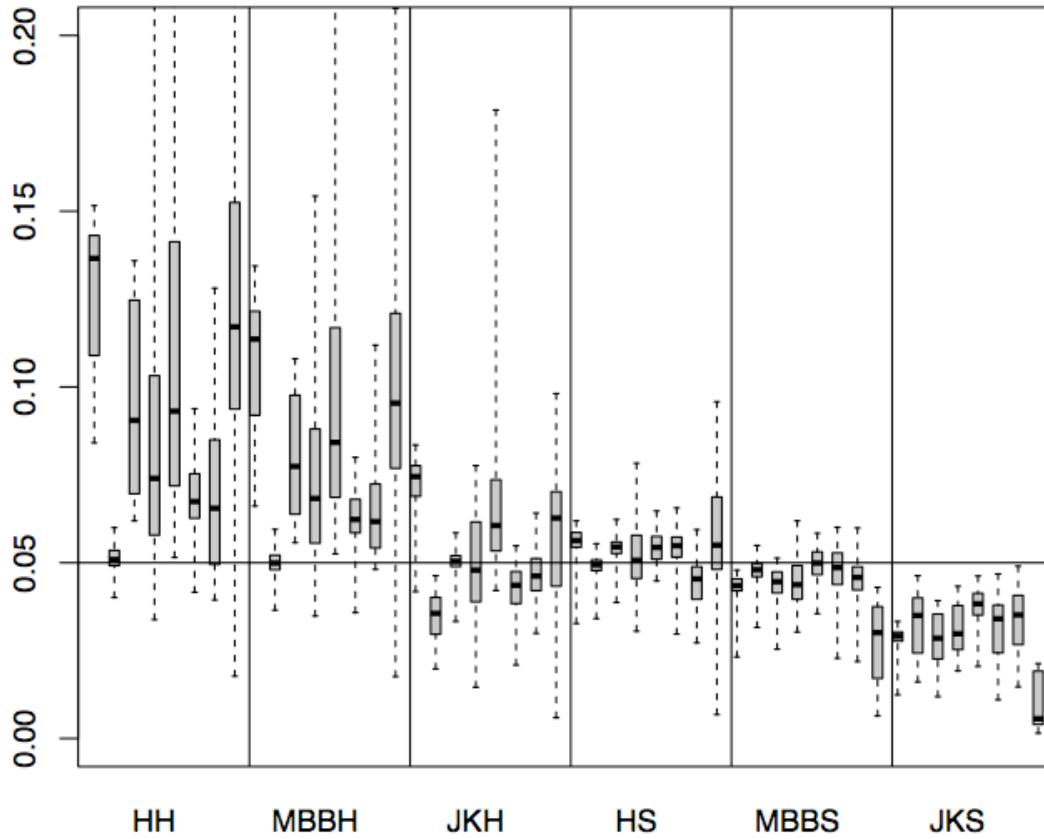
## Appendix B. Tables and Figures

**Table 1: Example analysis using RVE with HTJ, MBBS, and JKS small sample corrections**

Coefficient	B	SE(B) HTJ	SE(B) MBB/HTJ	SE(B) JK/HTJ	df MBBS	df JKS	Prob( t >) HTJ	Prob( t >) MBBS	Prob( t >) JKS
Intercept	0.704	0.320	1.044	1.252	9.7	7.9	0.030	0.055	0.102
Personal	0.177	0.104	1.087	1.359	3.1	2.5	0.103	0.223	0.313
% white	-0.950	0.389	1.051	1.267	4.3	3.0	0.023	0.089	0.150
Wave_c	-0.060	0.025	0.920	0.988	6.7	5.9	0.023	0.039	0.052
Wave_m	0.067	0.106	0.991	1.119	13.7	12.8	0.531	0.530	0.580

*Note: the above analysis uses approximately inverse-average variance weights based on an RVE model using correlated effects weights with an assumed  $\rho = 0.80$ . The between study variation used for these weights was  $\tau^2 = 0.037$ . The model is based on  $m=28$  studies with a total of  $N=300$  effect sizes.*

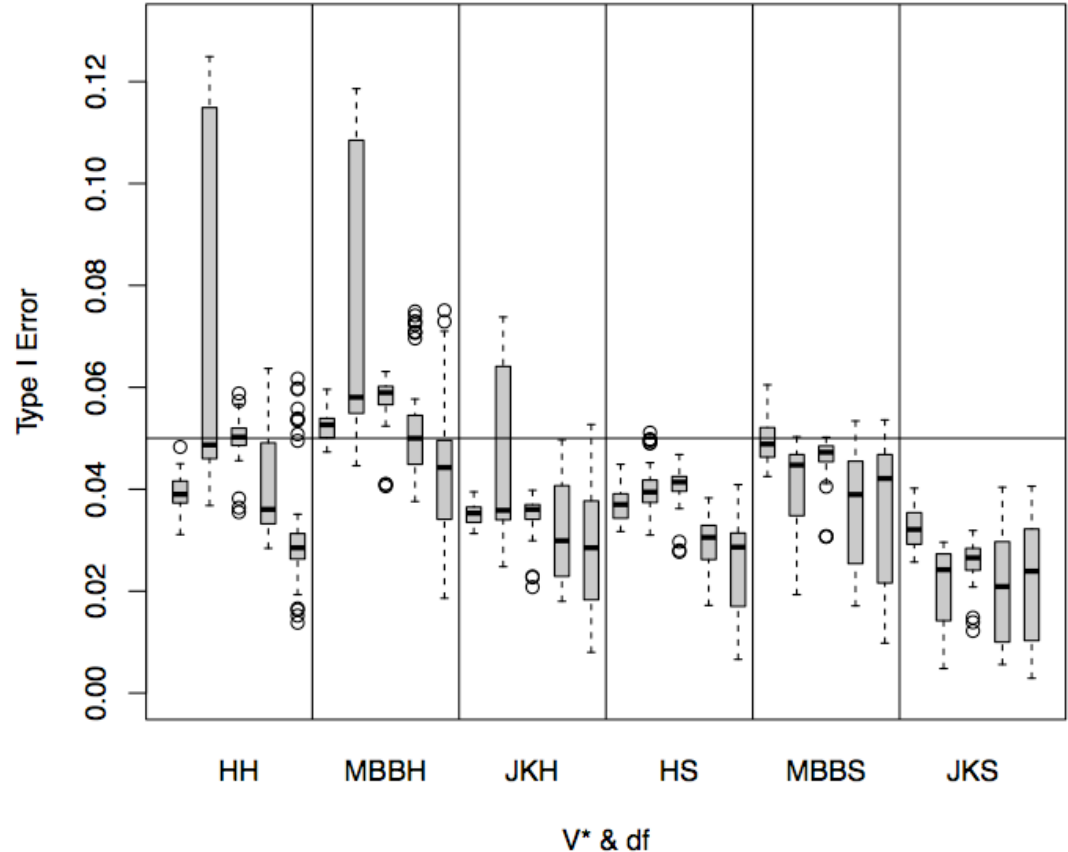
**Figure 1: Boxplot comparison of Type-I error rates of six variance estimators and eight variable types**



*Note: The first letters signify the adjustment, while the last letter signifies the degrees of freedom ( $H=m-2$ ,  $S$  = Satterthwaite); within each estimator, the variables from left to right follow those found in Table 3; for those with  $df=S$ , only values corresponding to  $df>4$  are shown.*

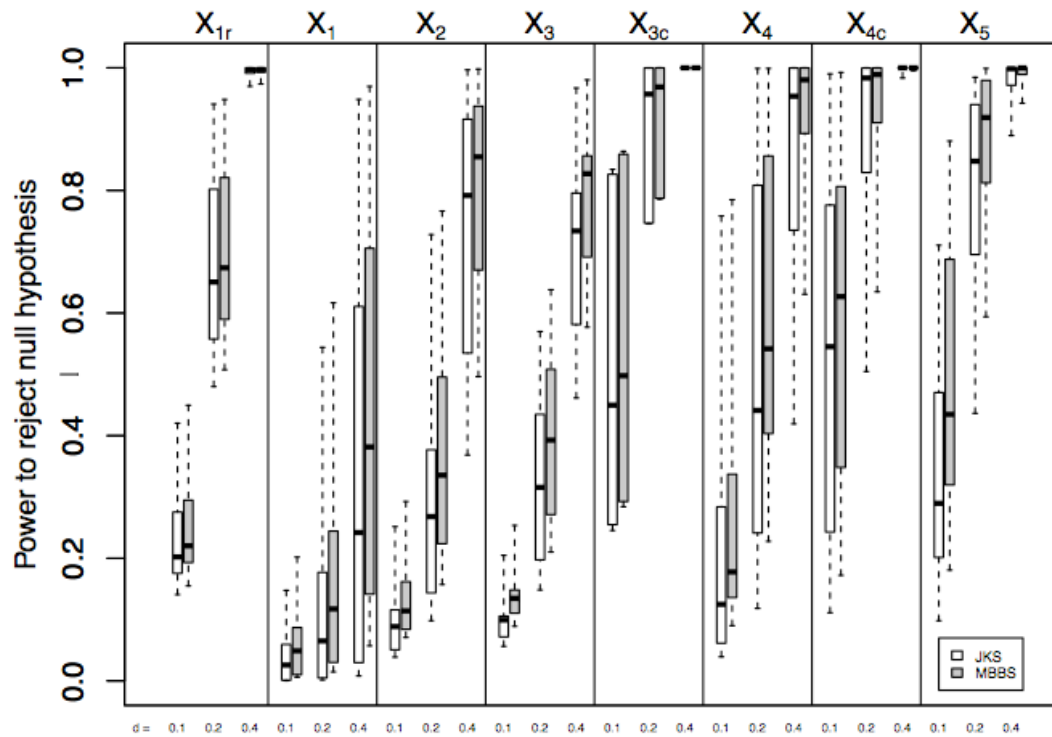


**Figure 2: Boxplot comparison of Type-I error rates of six variance estimators and meta-regression model with four variable**



Note: The first letters signify the adjustment, while the last letter signifies the degrees of freedom (H=m-p, S=Satterthwaite); within each estimator, the variables from left to right follow those found in Table 5; the results look across all three models and two values of k&n with m=20.

**Figure 3: Boxplot comparing the power of t-tests using MBBS and JKS adjustments**



Note: The bars indicate the power of the JKS test for each variable and effect size combination, across all parameter values studied. Only tests with Satterthwaite  $df > 4$  are shown.